DOCUMENT RESUME

ED 304 464 TM 012 836

AUTHOR Ellis, Michael V.; And Others
TITLE An Investigation of Supervision Research Methodology: Where Have We Gone Wrong?
PUB DATE Aug 88
NOTE 25p.; Paper presented at the Annual Meeting of the American Psychological Association (96th, Atlanta, GA, August 12-16, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Counseling; Effect Size; Literature Reviews; *Research Methodology; Research Problems; Research Reports; *Supervision; Validity
IDENTIFIERS *Counselor Supervisors; *Empirical Research; Error Analysis (Statistics); Power (Statistics)

ABSTRACT

The purpose of this study was to investigate the methodology of empirical counseling supervision articles published from 1981 to 1987. Seventy-five studies met the criteria for inclusion. Fifty-five articles included adequate information about at least one inferential statistical test to permit computing eta(sup 2), epsilon(sup 2), power, and alpha(sub EW). The remaining 20 articles did not provide sufficient information to either complete these computations or undergo quantitative analysis. Sample and population effect sizes--eta(sup2), epsilon(sup 2), and a priori and post hoc statistical power (1-beta)--were computed for every statistical test reported in sufficient detail. Experimentwise Type I and Type II error rates--alpha(sub Ew), beta(sub Ew)--were calculated for each study. The studies were critiqued using T. D. Cook and D. T. Campbell's (1979) 33 threats of validity plus the 12 methodological threats used by R. K. Russell, A. M. Crimmings, and R. W. Lent (1984). Although researchers have improved their designs in several ways, numerous potent methodological flaws continue to beset supervision research (e.g., measures with poor psychometric properties, inflated Type I and Type II error rates). Recommendations and implications for research, theory, and practice are discussed. Three tables summarize study data, and a bibliography of the 75 reviewed articles is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED304464

AN INVESTIGATION OF SUPERVISION RESEARCH METHODOLOGY: WHERE HAVE WE GONE WRONG?

Michael V. Ellis, Nicholas Ladany, Maxine Krengel, and Deborah Schult

Department of Counseling Psychology
University at Albany, State University of New York

Paper presented at the annual meeting of the American Psychological Association, August 1988, Atlanta, Georgia.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Michael V. Ellis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Correspondence regarding this paper should be addressed to Michael V. Ellis, Department of Counseling Psychology, ED 220, University at Albany, SUNY, 1400 Washington Avenue, Albany, New York 12222.

TM012836

An Investigation of Supervision Research Methodology: Where Have We Gone Wrong?

Abstract

The purpose of this study was to investigate the methodology of empirical counseling supervision articles published from 1981 to 1987. Seventy-five studies met the criteria for inclusion. Sample and population effect sizes (η^2, ϵ^2), and a priori and post hoc statistical power (1-ß) were computed for every statistical test reported in sufficient detail. Experimentwise Type I and Type II error rates (α_{EW}, β_{EW}) were calculated for each study. The studies were critiqued using Cook and Campbell's (1979) 33 threats of validity plus the 12 methodological threats used by Russell, Crimmings, and Lent (1984). Although researchers have improved their designs in several ways, numerous potent methodological flaws continue to beset supervision research (e.g., measures with poor psychometric properties, inflated Type I and Type II error rates). Recommendations and implications for research, theory, and practice were discussed.

An Investigation of Supervision Research Methodology: Where Have We Gone Wrong?

The primary reasons for conducting reviews of the extant literature in a field of inquiry are to assess substantive and methodological advances, verify and develop theory, synthesize knowledge, and generalize from a set of studies (Jackson, 1980). As Feldman (1971) and others (e.g., Cooper, 1982; Green & Hall, 1984; Light & Smith, 1971; Rosenthal & Rubin, 1979, 1982; Taveggia, 1974) have argued, a good review "shows how much is known in an area, [and] also shows how little is known" (Feldman, 1971, p. 100). To be able to determine what is known, it is incumbent on the reviewer to ascertain if the variability among the studies reviewed is "due to sampling error, differences in the methodological adequacy of the studies, or differences in the phenomena that were studied" (Jackson, 1980, p. 447). Thus, the reviewer would need an in depth evaluation of the methodological and statistical (i.e., sampling error) components of the studies. One could first take into account the weaknesses in the methods of the studies when making inferences from the studies. Second, specific recommendations could be proposed to advance the field of inquiry. With this in mind, the purpose of the study reported here was to investigate systematically the methodology of the data based literature in counseling supervision published from 1981 to 1987. The defintion of supervision used in this study was adapted from Loganbill, Hardy, & Delworth (1982). Supervision was defined as an intensive interpersonally focused relationship in which one person is designated to facilitate the development of therapeutic competence in the other person(s) (cf. Loganbill et al., 1982, p. 4).

In recent years there has been a proliferation of articles reviewing the theoretical and empirical manuscripts related to counseling supervision. Some reviewers seemed to turn a non-critical eye towards the methodology of the empirical supervision literature (e.g., Kaplan, 1983; Holloway, 1984; Lambert & Arnold, 1987; Leddick & Bernard, 1980; Liddle & Halpin, 1978; Matarazzo & Patterson, 1986). Other authors have examined the methodology and research designs as part of a critical review of the supervision research (e.g., Hansen, Pound, & Petro, 1976; Hansen, Robins, & Grimes, 1982; Hansen & Warner, 1971; Holloway & Johnston, 1985; Kurtz, Marshall, & Banspach, 1985; Loganbill et al., 1982; Worthington, 1987). Only two review articles presented details of the methodological flaws encountered in the studies reviewed (Holloway, 1987; Russell, Crimmings, & Lent, 1984). On closer scrutiny, however, Russell et al. appeared to be the sole reviewers to systematically rate the methodology of each study.

In their chapter, Russell et al. reviewed both the theoretical and empirical literature on one-on-one supervision published prior to 1981. Their critique of the research included rating each study in terms of six threats to internal validity (e.g., no randomization, inadequate sample size) and six threats to the external validity (e.g., nonrepresentative supervisee or supervisor population, overly brief training period). Yet, the specifics of the rating criteria, of the raters, or of interrater reliability were not reported.

Given that the reviews published since Russell et al. have not included an in depth evaluation of the research methodology of the studies, it seemed that such an investigation would be valuable. Indeed, one of Kaplan's (1983) concluding remarks for his review were "perhaps the most productive focus of future literature on practicum supervision would be on research methodology" (p. 224). To this end, the methodology of research articles on supervision published from 1981 to 1987 were examined. A study was considered a supervision article if it was a data based article which pertained to individual, group, and peer supervision of counselor/ psychotherapy trainees. In order to address the methodology and sampling error issues, a qualitative and a quantitative analysis, respectively, were performed (see Haase, 1974; Jackson, 1980; Light & Smith, 1971; Rosenthal & Rubin, 1979).

Method

Data collection

The research studies reviewed herein were identified through four sources of information and selected for inclusion on the basis of several criteria. Four information retrieval approaches served to identify the 89 potential supervision manuscripts published from 1981 to 1987 (a) the ancestry approach (using reference lists from reviews and original supervision articles), (b) Psychological Abstracts, (c) an on-line computer search of the social science citations and the psychological abstracts data bases, and (d) systematically searching periodicals publishing supervision literature (i.e., The Clinical Supervisor, Counselor Education and Supervision, The Counseling Psychologist, Journal of Clinical Psychology, Journal of Counseling and Development / Personnel and Guidance Journal, Journal of Counseling Psychology, Professional Psychology: Research and Practice, and Psychotherapy: Theory, Research, and Practice). To be included a manuscript had to (a) meet the definition of supervision stipulated above, (b) be a data based article published between 1981 and 1987, inclusively, and (c) address supervision as part of the research question, methods, or use supervisors or supervisees as participants. An article was excluded if it involved either counselor training not specifically addressing supervision, or speech pathology, or training of group therapists (see Gazda, 1984).

Quantitative measures

For each inferential statistic reported in a study, several measures were calculated for the quantitative analysis. The particular measures were selected based on previous quantitative reviews (Cohen, 1962; Haase, 1974; Haase, Waechter, & Solomon, 1982). Strength of association (i.e., size or magnitude of the effect) was measured by eta squared (η^2), a sample or descriptive statistic representing the proportion of dependent variable (DV) variance accounted for by the independent variable (IV). Eta squared (η^2) was computed using a formula that employs the value of the test statistic (e.g., t, F) and relevant degrees of freedom (Haase, 1983). Consequently, some of the values of η^2 reported herein are technically values of partial η^2 rather than classical η^2 (see Kennedy, 1970). Partial η^2 values are slightly higher than values of classical η^2; however, it is expected that the overestimate of sample effect size should not have serious consequences for the other measures computed using η^2 (see Haase et al.,

1982). The value of ϵ^2 (epsilon squared) was computed as a population estimate of the strength of association (effect size) using Ezekiel's (1930) shrinkage formula. Although a negatively biased estimate of the population parameter, the bias of ϵ^2 should be less than .1/N (Montgomery & Morrison, 1973).

Statistical power is the conditional probability of detecting an effect given that the effect exists in the population (i.e., the probability of rejecting a false null hypothesis). Post hoc power of the statistical tests was based on sample η^2 whereas a priori power was estimated using ϵ^2, the population effect size (Cohen, 1977). Both probabilities were computed using Abramowitz and Stegun's (1972) procedure and Cohen's (1977) formulas. The exact probability (p) of the statistical test (e.g., F, t) was derived using a procedure presented by Jaspen (1965).

For each study, the experimentwise Type I error rate (α_{EW}) was computed using $\alpha_{EW} = 1 - (1 - \alpha_{PC})^S$, where α_{PC} is the per comparison alpha (typically .05) and s is the number of separate analyses performed in the study. The following criteria were used to determine the number of separate analyses performed in a study. Following Fisher's least significant difference procedure, post hoc or follow-up tests (e.g., a protected univariate F test following a significant multivariate test) were not counted as separate entities. All other analyses were treated as separate entities. Hence, each main effect and each interaction test of multifactor Anova (or Manova) designs were considered separately.

Qualitative Measures

For the qualitative analysis, the methodology of each study was evaluated in terms of potential threats to the validity of the results. Cook and Campbell (1979) proposed four classes of validity as well as the threats to each: statistical conclusion validity (7 specific threats), internal validity (13 threats), construct validity (10 threats), and external validity (3 threats). For each of the 33 threats, the study was rated as yes (definitely a threat), no (not a threat), or not enough information to evaluate (if the threat applied or not). The 12 methodological deficiencies employed by Russell et al. (1984) were also used to critique the studies, thus permitting comparisons to their earlier review. Each study was critiqued further according to the methodological principles from Kazdin (1986) and Kerlinger (1986). The most potent methodological issues that could explain or limit the results were subsequently listed.

One counseling psychologist and three advanced doctoral students in counseling psychology served as raters. Raters were trained for 10 hours in the use of the rating procedures. Every article was rated twice by two judges who worked in the same two pairs. Differences in ratings were resolved by consensus. Across all 45 rating criteria (threats), the average interrater agreements (unanimous) was 90.25% (SD = 7.31). Only four criteria had less than 80% perfect agreement: heterogeneity of participants (74.0%), confounding constructs with limited levels of constructs (75.0%), restricted generalizability across constructs (71.0%), and restricted range of dependent variables (70.0%).

Results

Of the 75 articles critiqued, 55 included adequate information about at least one inferential statistical test to permit computing η^2, ϵ^2, power, and α_{EW}. Fifteen of the remaining 20 articles did not provide sufficient information to complete the computations for this analysis (i.e., usually lacking the degrees of freedom; the specific values for F or t; or the exact number of analyses performed). Five studies were survey research designs relying on descriptive statistics (see Kerlinger, 1986), and thus were not appropriate for the quantitative analysis.

The results of the quantitative analysis are presented in Table 1. Examining Table 1 revealed that the median sample size (per statistical test) was 49.0. The median sample effect size (η^2) across the 1581 statistical tests was .061 and the median population effect size (ϵ^2) was .031. Hence, the average sample variance of the dependent variable (DV) accounted for by the independent variable (IV) was 6.1% which translated to 3.1% after shrinkage (controlling for sample size and number of treatment groups). The median post hoc power ($1 - \beta_{PC}$) was .403 whereas the median a priori power was .231. Given the conventionally desired level of power of .80 (Cohen, 1977), the percentiles corresponding to .80 for post hoc power and for a priori power were the 30th and 25th, respectively. The median observed per comparison Type II error rate (β_{PC}) was .597 (i.e., the probability of failing to reject a false null hypothesis). The best estimate of the experimentwise Type I error rate across the 55 studies was a median α_{EW} = .750. Using the median β_{PC} = .597 and a median of 13.5 analysis per study, the estimated experimentwise Type II error rate (β_{EW}) was .9999 (see Haase & Ellis, 1987, for the formula to calculate β_{EW}).

The results of the qualitative analyses are summarized in Table 2. Several specific findings were notable. Assaying statistical conclusion validity, over 80% of the 75 studies were judged to suffer from inflated Type I or Type II error rates and unreliability of the dependent or independent measures. Regarding internal validity, selection and ambiguity of causal direction were threats to more than 70% of the studies. On the positive side, more than 80% of the studies were found to be free of 9 of the 13 threats to internal validity. Three threats to construct validity were found in 68% or greater of the studies: mono-method bias (81.3%), confounding the construct with limited levels of the construct (76%), and inadequate preoperationalization explication of the constructs (68%). All three of the threats to external validity plagued in excess of 95% of the studies critiqued. Applying Russell et al.'s (1984) 12 threats resulted in finding 76% of the studies suffering from inadequate sample size and 68% relied exclusively on self-report data.

After making the 45 ratings, the most potent (3 to 5) methodological threats for each study were listed. The 17 most frequently occurring of these threats (29 total) that pertained to at least 10% of the studies are presented in Table 3. As can be seen in Table 3, 7 of the 17 most salient threats were to the statistical conclusion validity; 4 to the internal validity, 4 to the construct validity, and 4 to the external validity of the studies, although not exclusively.

Discussion

The purpose of this study was to assess the methodological strengths and limitations of the population of 75 data based supervision articles published from 1981 to 1987. To accomplish this task, we examined both quantitative and qualitative data about the published studies. The quantitative results (i.e., effect size, statistical power, and experimentwise error rates) were comparable to results reported from earlier reviews (Cohen, 1962; Haase, 1974; Haase et al., 1982). The median effect size found here (η^2 = .061) is virtually the same as Cohen's (1977) medium effect size (η^2 = .0588) and somewhat lower than Haase et al.'s (1982) median effect size (η^2 = .083). The difference in Mdn effect size between Haase et al. and our study could be attributable to using measures with poor psychometric properties (i.e., reliability, validity). The statistical power (Mdn = .403) was quite similar to that found previously for the counselor education literature (1-ß = .365; Haase, 1974) and the abnormal-social psychology literature (1-ß = .460; Cohen, 1962). These values were far from the conventionally desired level for power of .7 or .8 (Cohen, 1977). On the positive side, at least 30% of the statistical analyses did have adequate power (i.e., greater than .80). As would be expected, the estimated population effect size (Mdn = .031) and statistical power based on this value (Mdn = .231) were substantially lower than the sample values. In terms of median experimentwise Type I and estimated experimentwise Type I error rates, both were considerably larger (α_{EW} = .750, $ß_{EW}$ = .9999) than similar reported estimated values (α_{EW} = .37, $ß_{EW}$ = .87; Haase & Ellis, 1987, p. 404). In fact only 5.5% of the studies had an experimentwise alpha less than or equal to 0.1. The discrepancy between these sets of experimentwise alpha and experimentwise beta appeared to be due to the number of supervision studies reviewed here that performed a large number of separate statistical tests with no or minimal attempts to control Type I or Type II error rates (e.g., Heppner & Roehlke, 1984; Worthington, 1984). It seems puzzling that an investigator would be comfortable conducting a study with more than a fifty percent chance of obtaining at least one spurious significant result (Type I error); yet, such was the case for over 70% of the studies reviewed here.

In short, the quantitative results for the last 8 years of supervision research were fairly typical to other empirical literatures. Researchers can expect to encounter medium effect sizes when investigating supervision phenomenon. A more unpleasant picture emerges regarding the experimentwise Type I and Type II error rates and statistical power. The majority of the investigations of supervision were simultaneously both unlikely to detect true effects and very likely to yield spurious significant results. We are left to conclude that the suggestions posited by Haase (1974) have gone unheeded by researchers in counselor education and supervision.

In terms of the qualitative analyses, many specific findings merit comment. Our discussion, however, will be limited to the more important findings. The interested reader is encouraged to examine the results (Tables 2 and 3) in greater detail. On the positive side, 10 of the 13 threats to internal validity and 6 of the 10 threats to construct validity were either controlled or not applicable to at least 76% of the studies (see Table 2). The reader should keep in mind that the first 5 threats to internal validity were specific to pre-post designs (repeated

measures) which were infrequent among the 75 studies. Of the 12 methodological threats used by Russell et al. (1984), fully two-thirds (8) showed marked improvement from the empirical supervision literature published before 1981 compared to the studies published after 1980. For example, there were 29% fewer studies surveyed in this review that used nonrepresentative supervisees or supervisors (from 40% to 10.7%). Hence, researchers have responded to many suggestions offered by Russell et al. and others (e.g., Hansen et al., 1979, 1982).

On the bleaker side, many of the studies reviewed herein were beset with methodological problems and threats to the validity of the results (cf. Tables 2 and 3). Unfortunately, 4 methodological threats (3 internal validity, 1 external validity) were found more frequently in our set of studies than Russell et al.'s (1984) review. Upon closer examination of the studies reported in Russell et al., the increase in these 4 methodological threats appeared to result from a shift in researchers using experimental analogue designs (pre 1981) to relying on quasi-experimental and ex post facto designs using real supervisors and supervisees (post 1980).

The appearance of the top six most salient methodological threats is noteworthy (see Table 3), yet not unexpected based on previous reviews of the supervision literature (e.g., Holloway, 1987; Russell et al., 1984). A striking and somewhat alarming finding was that in fully 26% of the studies the assumptions underlying the statistics used were violated to the extent that the validity of the results was severely compromised. The most common violations were due to (a) heterogeneity of variance, often attributable to widely discrepant cell sizes, (b) the misapplication or inappropriate use of a statistical procedure (e.g., using repeated measures ANOVA with independent data), and (c) the failure to meet the "absolute" minimum 5:1 ratio of the number of participants to the number of items/variables for a factor analysis or principal components analysis (see Tinsley & Tinsley, 1976).

The results of this study are not without limitations. Perhaps the most important aspect concerns the 45 criteria used to rate the studies. Several criteria had no or little variability (e.g., instrumentation, statistical regression, diffusion of treatment, compensatory equalization, rivalry by participants, resentful demoralization, the three external validity threats, and overly brief training period); thus suggesting that these particular criteria were not applicable to the last 8 years of supervision research. It could also be the case that the criteria were defined so as to preclude detecting variations in the designs currently employed in supervision research (we adhered closely to Cook & Campbell's definitions). Future reviewers may want to utilize or develop other rating criteria better suited to supervision research. Another limitation resides in the 5 criteria that had interrater agreements in the .70s (see Table 2). A fairly large percent of studies (greater than 50%) were judged as being susceptible to 4 of these threats. Hence, the results for the 5 criteria with lower reliability should be treated with caution.

Many studies failed to provide complete information on all the statistical tests performed, obviating the computation of quantitative data. In fact, the exact number of analyses performed that yielded nonsignificant results was frequently impossible to determine. These, in combination with the likelihood that studies

with nonsignificant results fail to be published, make the results of the quantitative analyses very liberal (optimistic). Hence, we would expect the values of the median effect size and median power to actually be smaller and the values of the experimentwise error rates to be larger. In other words, a gloomier picture emerges.

Implications

An incredible amount of resources, both human and financial, were required to conduct the research critiqued in this review. The efforts of these researchers has expanded our understanding of the complexities and vicissitudes of counselor supervision. We expect that researchers will continue in their endeavors to improve the methods and research designs used in investigations of supervision. In particular, we exhort future researchers in supervision to attend more closely to the threats to statistical conclusion and internal validity so that their research will not be beset with the problems found in the supervision research over the last 8 years. To this end, the following recommendations are made.

Statistical Conclusion Validity

1. Measures of dependent and independent variables

   - There continues to exist a need for psychometrically sound measures for supervision (Holloway, 1984, 1987; Lambert & Arnold, 1987). These measures should be theory driven and validated to meet the specified purposes rather than adhere to the trinitarian approach to test construction and validation (Landy, 1986). Use state-of-the-art procedures (Dawis, 1987; Jackson, 1970), such as confirmatory factor analytic methods (Long, 1983).

   - When adapting an existing measure designed for counseling settings to a supervision setting, the researcher should conduct minimally a pilot study to establish the psychometric properties of the measure in the new setting. The researcher is also encouraged to report the internal consistency (e.g., Chronbach's alpha) based on the full sample. Check for restriction in range on each scale and make changes if necessary (see Dawis, 1987).

   - Unless an expressed purpose of the study, avoid treating each item as a separate dependent measure, especially if they were factor analyzed. In this case, use factor scores rather than analyzing items separately in factorial clusters (e.g., Worthington & Stern, 1985).

   - When measures yield continuously distributed scores, analyze them as such. This avoids the 36% loss of information and statistical power that results from dichotomizing (or mutlichotomizing) the continuous variable (see Pedhauser, 1982, for regression approaches to ANOVA type problems).

   - When using raters and rating procedures, provide sufficient information about the raters and their training for the reader to evaluate the raters' qualifications, potential for biases (e.g., blind to experimental conditions), standardization of ratings, and how rating materials were assigned to raters.

2. Type I and Type II error rates

   - Be cognizant of error rate issues in addition to methods to control escalating error rates, for example, omnibus multivariate analyses, Bonferroni adjusted alpha levels, and a priori power analyses (Bray & Maxwell, 1982; Cohen, 1977; Haase et al., 1982; Haase & Ellis, 1987). Avoid fishing expeditions; judiciously choose statistical procedures that match the design/methods to test the hypotheses. Select only those independent and dependent variables that are clearly linked to the constructs under investigation and for which adequate rationale has been provided.

3. Statistical procedures and assumptions

   - Researchers need to be familiar with the assumptions underlying the statistics being used. The assumptions of any statistic need to be assessed for violations. For example, in situations of disproportional cell sizes, a test of homogeneity of variance is warranted (e.g., Bartlett-Box *F* test, Box's *M*) and if necessary, make corrections (e.g., data transformation) or adjust the statistic (e.g., Jaccard & Ackerman, 1985).

   - Except in rare circumstances, confirmatory factor analytic procedures represent the preferred factor analytic paradigm because they utilize more objective criteria and force the researcher to explicate previously implicit models or theories (Long, 1983).

Internal Validity

1. Testing developmental models of supervision

   - The shortcomings of cross sectional designs have been well covered recently (Holloway, 1987). In short, given the limited utility of cross sectional designs, sequential designs are preferred. The sequential designs allow for the investigation of intraindividual change and eliminates the threats of cohort effects by combining aspects of both cross-sectional and longitudinal designs (e.g., Bates, Reese, & Nesselroade, 1972).

2. Controlling rival hypotheses and confounds

   - Randomization (assignment to treatment or conditions) is the method of choice to control extraneous variables and error variance (Kerlinger, 1986). Blocking, matching then randomizing, using a covariate, homogenizing the sample, or treating a variable as an additional independent variable may be more practical alternatives given the nature of supervision research.

   - In cases of nonrandomized studies, full reporting of participant demographics (e.g., age, sex, degree program) and other data germane to the study (e.g., amount of supervised counseling experience, theoretical orientation, types of clients seen) are needed to assess for difference among groups. The specific variables/data selected to test for differences

should be used to rule out rival hypotheses (see Cook & Campbell, 1979). In fact, Cook and Campbell strongly recommend pretesting on salient variables to assess and eventually control pre-treatment group difference.

- When using treatments or manipulations, conduct pilot studies to substantiate that the manipulations work as predicted. Pilot studies can also be a means to maximize effect sizes before running the study proper. Perform validity or manipulation checks as the study progresses to avoid experimenter drift and maintain reliability of treatment implementation (i.e., treatment integrity). Include follow-up assessments as part of the design.

- Avoid using participants in multiple roles (e.g., as both supervisor in one group and supervisee in another). If this is unavoidable, control for this confound methodologically or statistically. Each investigation and design has confounds unique to the particular study. It is incumbent on the researcher to identify and control as many of these confounds as possible, or, note the limitations and interpret the results with the limitations clearly in mind.

- Keep experimenters, confederates, and judges blind to the experimental hypotheses under investigation to maintain procedural integrity.

Construct Validity

Theory, rationale, and definitions

- Ample models and theories of supervision exist to permit "theory driven" research. Researchers are strongly encouraged to test rigorously the models of supervision.

- Clearly define the constructs central to the study; this can help prevent equivocating terms (i.e., preoperational explication of the constructs).

- Make the study theory driven. Provide a lucid rationale that logically relates the constructs to one another. That is, make the implicit models and "theories" explicit.

- The research question, rationale, hypotheses, design and methods, operational definitions, and statistical procedures should all be matched logically and consistent with each other (see Cook & Campbell, 1979; Kerlinger, 1986).

- When interpreting and discussing the results, keep within the scope of the study and its limitations.

External Validity

- One of the best ways of establishing external validity is by replicating the study with an independent sample in a different setting using alternative operationalizations of the central constructs. Given the inherent difficulties in publishing replications, the multiple study article appears the avenue of choice. This will not only encourage programmatic research (replication with extensions), but also provide evidence of external validity (see Friedlander & Ward, 1984).

- Whenever possible use bonafide supervisors, supervisees, and clients as participants and confederates. Keep the method as realistic to actual supervision situations as possible.

- Use multiple methods of data collection rather than relying on self-report data.

- Whenever possible, randomly sample from a target population. Use stratified random samples when appropriate to the research design and hypotheses.

Survey Research

- An extensive literature exists on designing and conducting quality survey research research (e.g., Kerlinger, 1986; Warwick & Lininger, 1975). Investigators attempting to conduct survey research should follow the guidelines outlined in the references cited above. For example, the survey instrument should be constructed carefully by piloting the instrument during development to ascertain its validity and reliability.

Consumers of Research

- Avoid accepting the results and discussion sections of a study at face value. Read with a critical eye, especially towards the statistical conclusion and internal validities of the study. Look for unrecognized flaws and confounds as well as strengths of the study.

- The profound Type I and Type II error rates in the supervision literature in combination with the near absence of replications suggest that the use or application of the majority of results from the literature (especially the last 8 years) is premature.

References

Abramowitz, B., & Stegun, A. (1972). Handbook of mathematical functions. New York: Dover.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks-Cole.

Bray, J. H., & Maxwell, S. E. (1982); Analyzing and interpreting significant MANOVAs. Review of Educational Research, 52, 340-367.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Academic Press.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston: Houghton Mifflin.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Ezekiel, M. (1930). Methods of correlational analysis. New York: Wiley.

Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44, 86-102.

Gazda, G. M. (1984). Group counseling: A developmental approach (3rd ed.). Boston: Allyn & Bacon.

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. Annual Review of Psychology, 35, 37-53.

Haase, R. F. (1974). Power analysis of research in counselor education. Counselor Education and Supervision, 14, 124-132.

Haase, R. F. (1983). Classical and partial eta square in multifactor Anova designs. Educational and Psychological Measurement, 43, 35-39.

Haase, R. F., & Ellis, M. V. (1987). Multivariate analysis of variance. Journal of Counseling Psychology, 34, 404-413.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. Journal of Counseling Psychology, 29, 58-65.

Hansen, J. C., Pound, R., & Petro, C. (1976). Review of research on practicum supervision. Counselor Education and Supervision, 16, 107-116.

Hansen, J. C., Robins, T. H., & Grimes, J. (1982). Review of research on practicum supervision. Counselor Education and Supervision, 22, 15-24.

Hansen, J. C., & Warner, R. W., Jr. (1971). Review of research on practicum supervision. Counselor Education and Supervision, 10, 261-272.

Holloway, E. L. (1984). Outcome evaluation in supervision research. The Counseling Psychologist, 12(4), 167-174.

Holloway, E. L. (1987). Developmental models of supervision: Is it development? Professional Psychology: Research and Practice, 18, 209-216.

Holloway, E. L., & Johnston, R. (1985). Group supervision: Widely practiced but poorly understood. Counselor education and Supervision, 24, 332-340.

Jaccard, J., & Ackerman, L. (1985). Repeated measures analysis of means in clinical research. Journal of Consulting and Clinical Psychology, 53, 426-428.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Speilberger (Ed.), Current topics in clinical and community psychology (Vol. 2, pp. 61-96). Orlando, FL: Academic Press.

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Jaspen, N. (1965). The calculation of probabilities corresponding to values of z, t, F, and chi-square. Educational and Psychological Measurement, 25, 877-880.

Kaplan, D. M. (1983). Current trends in practicum supervision research. Counselor Education and Supervision, 22, 215-226.

Kazdin, A. E. (1986). Research designs and methodology. In S. L. Garfield & A. E. Bergin (Eds.), Handbook of psychotherapy and behavior change (3rd ed., pp. 23-68). New York: Wiley.

Kennedy, J. J. (1970). The eta coefficient in complex Anova designs. Educational and Psychological Measurement, 30, 885-889.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart, & Winston.

Kurtz, P. D., Marshall, E. K., & Banspach, S. W. (1985). Interpersonal skill-training research: A 12-year review and analysis. Counselor Education and Supervision, 24, 249-263.

Lambert, M. J., & Arnold, R. C. (1987). Research and the supervisory process. Professional Psychology: Research and Practice, 18, 217-224.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41, 1183-1192.

Leddick, G. R., & Bernard, J. M. (1980). The history of supervision: A critical review. Counselor Education and Supervision, 20, 187-196.

Liddle, H. A., & Halpin, R. J. (1978). Family therapy and supervision literature: A comparative review. Journal of Marriage and Family Counseling, 4, 77-98.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.

Loganbill, C., Hardy, E., & Delworth, U. (1982). Supervision: A conceptual model. The Counseling Psychologist, 10(1), 3-42.

Long, J. (1983). Confirmtory factor analysis. Beverly Hills, CA: Sage.

Matarazzo, R. G., & Patterson, D. (1986). Research on the teaching and learning of therapeutic skills. In S. L. Garfield & A. E. Bergin (Eds.), Handbook of psychotherapy and behavior change (3rd ed., pp. 821-843). New York: Wiley.

Montgomery, D. B., & Morrison, D. A. (1973). A note on adjusting R^2. Journal of Finance, 28, 1009-1013.

Pedhazur, E. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart & Winston.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. Psychological Bulletin, 86, 1165-1168.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.

Russell, R. K., Crimmings, A. M., & Lent, R. W. (1984). Counselor training and supervision: Theory and research. In S. D. Brown & R. W. Lent (Eds.), Handbook of counseling psychology (pp. 625-681). New York: Wiley.

Taveggia, T. C. (1974). Resolving research controversy through empirical cumulation: Toward reliable sociological knowledge. Sociological Methods and Research, 2, 395-407.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. Journal of Counseling Psychology,. 34, 414-424.

Warwich, D., & Lininger, C. (1975). The sample survey: Theory and practice. New York: McGraw-Hill.

Worthington, E. L., Jr. (1987). Changes in supervision as counselors and supervisors gain experience: A review. Professional Psychology: Research and Practice, 18, 189-208.

Reviewed Articles

Allen, G. J., Szollos, S. J., & Williams, B. E. (1986). Doctoral students' comparative evaluations of best and worst psychotherapy supervision. Professional Psychology: Research and Practice, 17, 91-99.

Barth, R. P., & Gambrill, E. D. (1984). Learning to interview: The quality of training opportunities. The Clinical Supervisor, 2, 3-13

Borders, L. D., Fong, M. L., & Neimeyer, G. J. (1986). Counseling students' level of ego development and perceptions of clients. Counselor Education and Supervision, 26, 36-49.

Carey, J. C., & Williams, K. S. (1986). Cognitive style in counselor education: A comparison of practicum supervisors and counselors in training. Counselor Education and Supervision, 26, 128-136.

Couchan, W. D., & Bernard, J. M. (1984). Effects of timing of supervision on supervisor and counselor performance. The Clinical Supervisor, 2, 3-20.

Claiborn, C. D., & Dixon, D. N. (1982). The acquisition of conceptual skills: An exploratory study. Counselor Education and Supervision, 21, 274-281.

Cross, D. G., & Brown, D. (1983). Counselor supervision as a function of trainee experience: Analysis of specific behaviors. Counselor Education and Supervision, 22, 333-341.

Dodenhoff, J. T. (1981). Interpersonal attraction and direct-indirect supervisor influence as predictors of counselor trainee effectiveness. Journal of Counseling Psychology, 28, 47-52.

Eisikovits, Z., Shurka, E., & Baizeman, M. (1983). Israeli social workers supervision preferences and practices: An example of a professional ideology. The Clinical Supervisior, 1, 35-42.

Ellis, M. V., & Dell, D. M. (1986). Dimensionality of supervisor roles: Supervisors' perceptions of supervision. Journal of Counseling Psychology, 33, 282-291.

Friedlander, M. L., Keller, K. E., Peca-Baker, T. A., & Olk, M. E. (1986). Effects of role conflict on counselor trainees' self-statements, anxiety level, and performance. Journal of Counseling Psychology, 33, 73-77.

Friedlander, M. L., & Snyder, J. (1983). Trainees' expectations for the supervisory process: Testing a developmental model. Counselor Education and Supervision, 22, 342-348.

Friedlander, M. L., & Ward, L. G. (1984). Development and validation of the supervisory styles inventory. Journal of Counseling Psychology, 31, 541-557.

Goodyear, R. K., Abadie, P. D., & Efros, F. (1984). Supervisory theory into practice: Differential perception of supervision by Ekstein, Ellis, Polster, and Rogers. Journal of Counseling Psychology, 31, 228-237.

Goodyear, R. K., & Robyak, J. E. (1982). Supervisors' theory and experience in supervisory focus. Psychological Reports, 51, 978.

Handley, P. (1982). Relationship between supervisors' and trainees' cognitive styles and the supervision process. Journal of Counseling Psychology, 29, 508-515.

Hart, G. M., & Falvey, E. (1987). Field supervision of counselors in training: A survey of the North Atlantic region. Counselor Education and Supervision, 26, 204-212.

Heppner, P. P., & Handley, P. (1982). The relationship between supervisory behaviors and perceived supervisor expertness, attractiveness, or trustworthiness. Counselor Education and Supervision, 22, 37-46.

Heppner, P. P. & Handley, P. G. (1981). A study of the interpersonal influence process in supervision. Journal of Counseling Psychology, 28, 437-444.

Heppner, P. P., & Roehlke, J. J. (1984). Differences among supervisees at different levels of training: Implications for a developmental model of supervision. Journal of Counseling Psychology, 31, 76-90.

Hess, A. K., & Hess, K. A. (1983). Psychotherapy supervision: A survey of internship training practices. Professional Psychology: Research and Practice, 14, 504-513.

Hill, C. E, Charles, D., & Reed, K. G. (1981). A longitudinal analysis of changes in counseling skills during doctoral training in counseling psychology. Journal of Counseling Psychology, 28, 428-436.

Holloway, E. L. (1982). Characteristics of the field practicum: A national survey. Counselor Education and Supervision, 22, 75-80.

Holloway, E. L. (1982). Interactional structure of the supervision interview. Journal of Counseling Psychology, 29, 309-317.

Holloway, E. L., & Wampold, B. E. (1983). Patterns of verbal behavior and judgments of satisfaction in the supervision interview. Journal of Counseling Psychology, 30, 227-234.

Holloway, E. L., & Wolleat, P. L. (1981). Style differences of beginning supervisors: An interactional analysis. Journal of Counseling Psychology, 28, 373-376.

Hutt, C. H., Scott, J., & King, M. (1983). A phenomenological study of supervisees' positive and negative experiences in supervision. Psychotherapy: Theory, Research, and Practice, 20, 118-123.

Kennard, B. D., Stewart, S. M., & Gluck, M. M. (1987). The supervision relationship: Variables contributing to positive versus negative experiences. Professional Psychology: Research and Practice, 18, 172-175.

Lanning, W. (1986). Development of the supervisor emphasis rating form. Counselor Education and Supervision, 25, 191-196.

Loizos, M. M., & Gold, J. (1981). Differential supervision and cognitive structure effects on empathy and counseling effectiveness. International Journal of Advanced Counseling, 4, 119-129.

Malikiosi-Loizos, M., Gold, J., Mehnert, W. O., & Work, G. G. (1981). Differential supervision and cognitive structure effects on empathy and counseling effectiveness. International Journal of Advanced Counseling, 4, 119-129.

Marikis, D. A., Russell, R. K., & Dell, D. M. (1985). Effects of supervisor experience level on planning and in-session supervisor verbal behavior. Journal of Counseling Psychology, 32, 410-416.

Martin, G. E., & McBride, M. C. (1987). The result of the implementation of a professional supervision model on counselor trainee behavior. Counselor Education and Supervision, 27, 2, 155-167.

McColley, S. H., & Baker, E. L. (1982). Training activities and styles of beginning supervisors: A survey. Professional Psychology, 13, 283-292.

McNeill, B. W., Stoltenberg, C. D., & Pierce, R. A. (1985). Supervisees' perceptions of their development: A test of the counselor complexity model. Journal of Counseling Psychology, 32, 630-633.

Miars, R. D., Tracey, T. J., Ray, P. B., Cornfeld, J. L., O'Farrell, M., & Gelso, C. J. (1983). Variation in supervision process across trainee experience levels. Journal of Counseling Psychology, 30, 403-412.

Morrison, J. K., Smith, F. J., & Becker, R. E. (1984). Differential impact of supervisor-subordinate transactions on ratings of community mental health team supervisors. The Clinical Supervisor, 2, 65-70.

Moy, C. T., & Goodman, E. D. (1984). A model for evaluating supervisory interactions in family therapy training. The Clinical Supervisor, 2, 21-29.

Petty, M. M., & Odewahn, C. A. (1983). Supervisory behavior and sex role stereotypes in human service organizations. The Clinical Supervisor, 1, 13-20.

Poertner, J., & Rapp, C. A. (1983). What is social work supervision? The Clinical Supervisor, 1, 53-65.

Rabinowitz, F. E., Heppner, P. P., & Roehlke, H. J. (1986). Descriptive study of process and outcome variables of supervision over time. Journal of Counseling Psychology, 33, 292-300.

Reising, G. N., & Daniels, M. H. (1983). A study of Hogan's model of counselor development and supervision. Journal of Counseling Psychology, 30, 235-244.

Rice, J. M., & Lutzker, J. R. (1983). Group and individual feedback, public posting, and prompting to increase counselor supervision. The Clinical Supervisor, 4, 77-90.

Richardson, B., & Stone, G. L. (1981). Effects of a cognitive adjunct procedure within a microtraining situation. Journal of Counseling Psychology, 28, 168-175.

Rickards, L. D. (1984). Verbal interaction and supervisor perception in counselor supervision. Journal of Counseling Psychology, 31, 262-265.

Robyak, J. E., Goodyear, R. K., & Prange, M. (1987). Effects of supervisor's sex, focus, and experience on preferences for interpersonal power bases. Counselor Education and Supervision, 26, 299-309.

Robyak, J. E., Goodyear, R. K., Prange, M. E., & Donham, G. (1986). Effects of gender, supervision, and presenting problems on practicum students' preference for interpersonal power bases. Journal of Counseling Psychology, 33, 159-163.

Rosenbaum, D. M. (1984). Evaluation of student performance in psycho-therapy. Journal of Clinical Psychology, 40, 1106-1111.

Rosenberg, E. H., Medini, G., & Lomranz, J. (1982). Factorial dimensions of supervisor-student evaluation. Journal of Social Psychology, 118, 105-111.

Rotholz, T., & Werk, A. (1984). Student supervision: An educational process. The Clinical Supervisor, 2, 15-27.

Russell, P. A., Lankford, M. W., & Grinnell, R. M. (1983). Attitudes toward supervisors in a human service agency. The Clinical Supervisor, 1, 57-71.

Sandell, R. (1985). Influence of supervision, therapist's competence and patient's ego level on the effects of time-limited psychotherapy. Psychotherapy and Psychosomatics, 44, 103-109.

Savickas, M. L., Marquart, C. D., & Supinski, C. R. (1986). Effective supervision in groups. Counselor Education and Supervision, 26, 17-25.

Schmidt, J. J., & Barret, R. L. (1983). Who's in charge? School counseling supervision in North Carolina. Counselor Education and Supervision, 23, 495-501.

Steinhelber, J., Patterson, V., Cliffe, K., & LeGoullon, M. (1984). An investigation of some relationships between psychotherapy supervision and patient change. Journal of Clinical Psychology, 40, 1346-1353.

Stenack, R. J., & Dye, H. A. (1982). Behavioral descriptions of counseling supervision roles. Counselor Education and Supervision, 21, 295-304.

Stenack, R. J., & Dye, H. A. (1983). Practicum supervision roles: Effects on supervisee statements. Counselor Education and Supervision, 23, 157-168.

Stoltenberg, C. D., Pierce, K. A., & McNeill, B. W. (1987). Effects of experience on counselor trainee needs. The Clinical Supervisor, 5, 23-32.

Stone, G. T. (1981). Effects of different strategies within a microtraining situation. Counselor Education and Supervision, 20, 301-311.

Tyler, J. D., & Weaver, S. H. (1981). Evaluating the clinical supervisee: A survey of practices in graduate training programs. Professional Psychology, 12, 434-437.

Ward, L. G., Friedlander, M. L., Schoen, L. G., & Klein, J. G. (1985). Strategic self-presentation in supervision. Journal of Counseling Psychology, 32, 111-118.

Wiley, M. O., & Ray, P. B. (1986). Counseling supervision by developmental level. Journal of Counseling Psychology, 33, 439-445.

Wilson, N. S., & Remley, T. P., Jr. (1987). Leadership in guidance: A survey of school counseling supervisors. Counselor Education and Supervision, 26, 213-220.

Worthington, E. L., Jr. (1984). Empirical investigation of supervision of counselors as they gain experience. Journal of Counseling Psychology, 31, 63-75.

Worthington, E. L., Jr. (1984). Use of trait labels in counseling supervision by experienced and inexperienced supervisors. Professional Psychology: Research and Practice, 14, 457-461.

Worthington, E. L., Jr., & Stern, A. (1985). Effects of supervisor and supervisee degree level and gender on the supervisory relationship. Journal of Counseling Psychology, 32, 252-262.

Yogev, S., & Pion, G. M. (1984). Do supervisors modify psychotherapy supervision according to supervisees' level of experience? Psychotherapy: Theory, Research and Practice, 21, 206-208.

Zucker, P. L., & Worthington, E. L., Jr. (1986). Supervision of interns and postdoctoral applicants for licensure in university counseling centers. Journal of Counseling Psychology, 33, 87-89.

Table 1

Means, Standard Deviations, and Quartiles for Quantitative Measures

			Percentile		
	M	SD	Mdn 50th	25th	75th
Across 1581 statistical tests					
N per analysis	63.55	73.39	49.00	33.00	73.00
Sample effect size (η^2)	.139	.183	.061	.015	.194
Population effect size (ϵ^2)	.117	.176	.031	.000	.163
Post hoc power	.488	.367	.403	.129	.897
A priori power	.408	.379	.231	.050	.795
Significance (p)	.258	.329	.089	.004	.544
Across 55 studies					
Experimentwise alpha (α_{EW})	.668	.303	.750	.370	.943
Corrected per comparison alpha (α_{PC})	.0024	.0024	.0016	.0007	.0032
Statistical tests per study (with sufficient information reported)	54.00	58.14	13.50	6.00	30.00

Table 2

Qualitative Analyses: Percentages per Categories and Interrater Reliabilities for 45 Methodological Threats Across 75 Supervision Studies

Threat	Category (%) Not a Threat	Insufficient Information	Definitely a Threat	Interrater Reliability[a]
Statistical Conclusion Validity				
low statistical power	41.3	4.0	54.7	81.0
violated assumptions of statistics	64.0	14.7	21.3	80.0
inflated error rate	9.3	1.3	89.3	95.0
unreliability of DV or IV measures	13.3	5.3	81.3	83.0
unreliability of treatment implementation	80.0	6.7	13.3	90.0
irrelevance in experimental setting	72.0	4.0	24.0	82.0
heterogeneity of participants	29.3	5.3	65.3	74.0
Internal Validity				
history	82.7	0.0	17.3	98.0
maturation	81.3	1.3	17.3	98.0
testing	81.3	1.3	17.3	93.0
instrumentation	93.3	2.7	4.0	93.0
statistical regression	94.7	1.3	4.0	95.0
selection	24.0	1.3	74.7	88.0
differential attrition	76.0	6.7	17.3	85.0
interactions with selection	58.7	0.0	41.3	90.0
ambiguity of causal direction	28.0	0.0	72.0	98.0
diffusion of treatment	98.7	0.0	1.3	100.0
compensatory equalization of treatments	98.7	0.0	1.3	99.0
rivalry by participants	98.7	0.0	1.3	99.0
resentful demoralization	100.0	0.0	0.0	100.0
Construct Validity				
inadequate preoperationalization explication	30.7	1.3	68.0	77.0
mono-operation bias	76.0	0.0	24.0	88.0
mono-method bias	18.7	0.0	81.3	87.0
hypothesis guessing within treatments	81.3	6.7	12.0	89.0
evaluation apprehension	85.3	4.0	10.7	86.0
experimenter expectancies	80.0	14.7	5.3	87.0
confounding construct with levels of construct	21.3	2.7	76.0	75.0
interaction of treatments	94.7	0.0	5.3	92.0
interaction of testing and treatments	76.0	0.0	24.0	90.0
restricted generalizability across constructs	48.0	0.0	52.0	71.0

(table continues)

Threat	Category (%) Not a Threat	Insufficient Information	Definitely a Threat	Interrater Reliability[a]
External Validity				
interaction of selection and treatment	0.0	1.3	98.7	98.0
interaction of setting and treatment	2.7	0.0	97.3	97.0
Interaction of history and treatment	5.3	0.0	94.7	98.0
Russell, Crimmings, & Lent's Threats				
lack of adequate control group	84.0	1.3	14.7	96.0
no pretreatment assessment	68.0	0.0	32.0	94.0
inadequate sample size	21.3	2.7	76.0	90.0
variations or confounds in length of training across conditions	93.3	2.7	4.0	91.0
nonrandom assignment to conditions	52.0	1.3	46.7	93.0
widely discrepant cell sizes	77.3	10.7	12.0	94.0
restricted range of DVs	33.3	38.7	28.0	70.0
nonrepresentative supervisee or supervisor population	88.0	1.3	10.7	91.0
lack of follow-up assessment	48.0	0.0	52.0	93.0
use of role play or audio-taped client to assess suprvised change	86.7	0.0	13.3	93.0
exclusive reliance on self-report data	32.0	0.0	68.0	94.0
overly brief training period	96.0	2.7	1.3	96.0

Note. Numbers represent the percent of the 75 studies assigned to the specific category for that methodological threat (criteria).

[a]Unanimous agreement, in percentages.

Table 3

Top 17 Most Salient Methodological Threats

Threat	Percent[a]	Type of Validity
Unreliability or invalidity of independent or dependent measures	72.0	s
Inflated Type I error rate	65.3	s
Nonrandom selection/nonrepresentative sample	53.3	e
Inflated Type II error rate/low statistical power	46.7	s
Nonrandom assignment	44.0	i
Cohort effects (cross sectional developmental designs)	28.0	i
Violated assumptions of statistics	26.7	s
Insufficient logic, rationale, or definitions	24.0	c
Not real supervisees, supervisors, or clients	21.3	e
Multichotomized a continuously distributed IV or DV	21.3	s
Differential ceiling/floor effects (restricted range) of DVs	20.0	i
Confounds: participants in multiple roles (supervisor and supervisee), supervisors with an unequal number of supervisees	17.3	i
Unreliability or invalidity of treatments/manipulations	16.0	s
No or inadequate participant demographic information	14.7	s,e
Experimenter expectancies or nonblind raters	10.7	c
Misinterpretation or overgeneralization of results	10.7	c,e
Mono-method bias	10.7	c

Note. [a]Percent of the 75 studies that the particular threat was judged a major flaw. s = statistical conclusion validity, i = internal validity, c = construct validity, e = external validity